Advances in Water Resources 116 (2018) 195–207

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres

Accounting for model error in Bayesian solutions to hydrogeophysical inverse problems using a local basis approach

Corinna Köpke[*,a], James Irving[a], Ahmed H. Elsheikh[b]

[a] University of Lausanne, Institute of Earth Sciences, Lausanne 1015, Switzerland
[b] Heriot-Watt University, School of Energy, Geoscience, Infrastructure and Society, Edinburgh EH14 1AS, United Kingdom

ARTICLE INFO

Keywords:
Model error
Bayesian inference
MCMC
Proxy model

ABSTRACT

Bayesian solutions to geophysical and hydrological inverse problems are dependent upon a forward model linking subsurface physical properties to measured data, which is typically assumed to be perfectly known in the inversion procedure. However, to make the stochastic solution of the inverse problem computationally tractable using methods such as Markov-chain-Monte-Carlo (MCMC), fast approximations of the forward model are commonly employed. This gives rise to model error, which has the potential to significantly bias posterior statistics if not properly accounted for. Here, we present a new methodology for dealing with the model error arising from the use of approximate forward solvers in Bayesian solutions to hydrogeophysical inverse problems. Our approach is geared towards the common case where this error cannot be (i) effectively characterized through some parametric statistical distribution; or (ii) estimated by interpolating between a small number of computed model-error realizations. To this end, we focus on identification and removal of the model-error component of the residual during MCMC using a projection-based approach, whereby the orthogonal basis employed for the projection is derived in each iteration from the K-nearest-neighboring entries in a model-error dictionary. The latter is constructed during the inversion and grows at a specified rate as the iterations proceed. We demonstrate the performance of our technique on the inversion of synthetic crosshole ground-penetrating radar travel-time data considering three different subsurface parameterizations of varying complexity. Synthetic data are generated using the eikonal equation, whereas a straight-ray forward model is assumed for their inversion. In each case, our developed approach enables us to remove posterior bias and obtain a more realistic characterization of uncertainty.

1. Introduction

Bayesian inversion of hydrological and geophysical data using Markov-chain-Monte-Carlo (MCMC) methods has become increasingly popular over the past decade. Key advantages of this approach are that: (i) it allows for more comprehensive quantification of posterior parameter uncertainty when compared to traditional linearized uncertainty estimates; (ii) it is extremely flexible in the sense that any information that can be expressed probabilistically (e.g., model prior information, data measurement errors) can be incorporated into the inverse problem; and (iii) it provides a natural framework within which to perform data integration. The Bayesian-MCMC approach does, however, have the notable disadvantage of being limited by its high computational cost, which results from the typically large numbers of model parameters in geophysical and hydrological problems combined with the need for small model perturbations along the Markov chain in order to ensure reasonable rates of proposal acceptance. That is, millions of forward model runs are commonly required to obtain meaningful posterior statistics, which is computationally prohibitive for many real-world applications (e.g., Ruggeri et al., 2015).

A variety of techniques exist for reducing the computational load of Bayesian-MCMC inversions. Recent algorithmic developments for MCMC methods, which take advantage of parallel architectures and incorporate chain history and posterior gradient information into the proposal distribution, have been shown to significantly improve computational efficiency past the standard Metropolis–Hastings approach (e.g., Haario et al., 2001; Marshall and Roberts, 2012; Neal, 2011; Sambridge, 2013; Stuart et al., 2004; Vrugt, 2016). Model reduction, through the use of basis functions that exploit the spatial correlation naturally present in subsurface properties (e.g., Davis and Li, 2011; Jafarpour et al., 2009; Linde and Vrugt, 2013; Oware et al., 2013), can also be performed to reduce the dimensionality, and thus the numerical complexity, of the inverse problem. Yet another means of reducing the computational load of Bayesian-MCMC inversions, and arguably the

* Corresponding author.
E-mail addresses: corinna.koepke@unil.ch (C. Köpke), james.irving@unil.ch (J. Irving), A.Elsheikh@hw.ac.uk (A.H. Elsheikh).

https://doi.org/10.1016/j.advwatres.2017.11.013
Received 21 August 2017; Received in revised form 2 November 2017; Accepted 12 November 2017
Available online 13 November 2017

most intuitive and commonly employed approach, is to use a fast approximation of the forward solver in place of the slower "full" numerical solution. This can be accomplished via simplification of the physics of the problem (e.g., Josset et al., 2015b; Scholer et al., 2012), reduction of the numerical accuracy of the solution by coarsening the model discretization (e.g., Arridge et al., 2006; Calvetti et al., 2014), or the construction of response-surface proxies based on, for example, polynomial chaos expansion, artificial neural networks, or Gaussian processes (e.g., Goh et al., 2013; Khu and Werner, 2003; Marzouk and Xiu, 2009; Rasmussen and Williams, 2006). While the use of approximate forward solvers in this manner can be highly effective, it can lead to strongly biased and overconfident posterior statistics if the discrepancies between the approximate and detailed solutions are not taken into account (e.g., Brynjarsdóttir and O'Hagan, 2014). Indeed, such "model errors" have the potential to overwhelm the effects of data measurement uncertainties and may have a controlling influence on posterior inference. Despite this fact, the issue of model error has been largely ignored in the vast majority of geophysical and hydrological studies to date where Bayesian-MCMC methods have been employed.

In recent years, a number of techniques have appeared in the scientific and engineering literature to address the model error problem, thus allowing for more effective use of approximate forward solvers in Bayesian stochastic inversions. One popular avenue of research focuses on the overall or "global" statistical characterization of these errors, whereby a small number of stochastic model-error realizations, generated by running the approximate and detailed forward solvers on random parameter sets drawn from the prior distribution, are used to develop likelihood functions that better reflect the combined nature of all error sources. To this end, by far the most straightforward and common approach is to assume that the model errors are Gaussian distributed and thus characterized by some mean vector and covariance matrix, both of which are estimated from the realizations (e.g., Arridge et al., 2006; Hansen et al., 2014; Kaipio and Somersalo, 2007; Lehikoinen et al., 2010; Stephen, 2007). Alternatively, customized parametric likelihood functions have been developed, most notably in the fields of catchment and urban hydrology, to reflect the non-Gaussian, strongly correlated, and often heteroscedastic nature of residuals in some problems (e.g., Del Giudice et al., 2013; Schoups and Vrugt, 2010; Smith et al., 2010, 2015). In all of these studies, it has been shown that inclusion of model-error statistical characteristics into the Bayesian likelihood function results in a broadening of posterior distributions along with, in many cases, a reduction in posterior bias. A key concern, however, is the validity of the assumption that the errors can be adequately described by the specified parametric distribution. Indeed, our own experience with high-dimensional spatially distributed inverse problems in geophysics and hydrology suggests that it is more often the case that model errors exhibit highly complex statistics and correlations that change significantly not only over the data space, but also as a function of the input model parameters. Note that this in part has led to greatly increased interest in alternative likelihood methods such as generalized likelihood uncertainty estimation (GLUE) (e.g., Beven and Binley, 1992) and approximate Bayesian computation (ABC) (e.g., Vrugt and Sadegh, 2013).

Another avenue of research to account for the discrepancy between approximate and detailed forward solvers in Bayesian stochastic inversions, which addresses the latter point above, focuses on the development of "local" error models that describe, either statistically or deterministically, the discrepancy between the approximate and detailed forward solutions over the model parameter space. O'Sullivan and Christie (2006), for example, use a small number of coarse-grid versus fine-grid model-error realizations, computed over a low-dimensional model-parameter space, to characterize through interpolation how the model-error mean and covariance matrix change as a function of the input parameters. Kennedy and O'Hagan (2001) present a comprehensive theoretical framework for dealing with model errors where the error statistics are described by a Gaussian process conditioned to the points in the parameter space where the model error is known. Xu and Valocchi (2015) also represent the model error as a Gaussian process that is trained during the Bayesian inversion with spatially and temporally distributed observations. Doherty and Christensen (2011) and Josset et al. (2015b) propose the use of regression models to predict the results of the detailed solver from the approximate solution, with the latter study making use of functional principal components analysis and dimension reduction to facilitate the analysis. Finally, Cui et al. (2011) assume that the model error obtained from the last detailed forward simulation during two-stage MCMC (discussed below) is a valid approximation of the model error for the current set of input parameters, and use it to correct the approximate solution before computing the likelihood. In all of this work, local error models are effectively constructed by interpolating between a limited number of model-error realizations, under the implicit assumptions that the model response surface is smooth enough to do so and that the parameter space has been adequately sampled. While this may be perfectly valid for low-dimensional inverse problems, it becomes extremely difficult in high dimensions.

Yet another means of addressing the issue of model error when using approximate forward solvers in Bayesian stochastic inversions is the two-stage MCMC approach. With this method, model errors are not explicitly accounted for, but instead are avoided altogether because the approximate solver is used only in a first accept/reject stage to prevent unpromising sets of model parameters from being tested with the computationally expensive detailed solution (e.g., Christen and Fox, 2005; Efendiev et al., 2009; Laloy et al., 2013; Ma et al., 2008). In order to realize computational gains with this technique, the approximate solver needs to be a "good" approximation in the sense that it provides results that are relatively close to the detailed one (Christen and Fox, 2005). For this reason, a number of researchers have paired the approximate solver with a local error model to improve its accuracy (Cui et al., 2011; Josset et al., 2015a; Laloy et al., 2013). The advantage of two-stage MCMC is that the effects of model errors in the Bayesian posterior distribution can be avoided. The significant disadvantage, however, is that the computational gains of the approach may still not be enough to render the inverse problem computationally tractable since each posterior realization must still pass through the detailed forward solver, in addition to other parameter sets that have passed the first stage but are later rejected.

In this paper, we attempt to address the above-mentioned challenges and present a new methodology for dealing with the model error arising from the use of approximate forward solvers in Bayesian solutions to hydrogeophysical inverse problems. Our approach is geared towards the common case where this error cannot be effectively characterized globally through some parametric statistical distribution or locally based on interpolation between a small number of computed realizations. Rather than focusing on the construction of a global or local error model, we instead work towards identification of the model-error component of the residual through a projection-based approach. In this regard, pairs of approximate and detailed model runs are stored in a dictionary that grows at a specified rate during the MCMC inversion procedure. At each iteration, a local model-error basis is constructed for the current test set of model parameters using the K-nearest neighbor (KNN) entries in the dictionary, which is then used to separate the model error from the other error sources. We begin in Section 2 with a brief review of Bayesian-MCMC methods followed by development of our modified approach to account for model error. We then show in Section 3 the application of our methodology to three example inversions involving crosshole ground-penetrating radar (GPR) traveltime tomography, where in each case different subsurface model parameterizations apply. In each example, posterior parameter distributions are compared for the cases where: (i) there is no model error present; (ii) model error is present but not accounted for; and (iii) model error is accounted for using our developed approach.